European University Institute
Economics Department

EUI Working Paper **ECO** No. 97/14

Mathematical and Statistical Modelling of Cointegration

Søren Johansen
European University Institute, Florence

========

Printed in Italy in April 1997
European University Institute
Badia Fiesolana
I – 50016 San Domenico (FI)
Italy

Mathematical and Statistical Modelling of Cointegration

Søren Johansen

European University Institute

March 12, 1997

Abstract

A brief overview is given of some of the problems in econometric model building. We discuss, by example, the concept of a time series, a random walk, and integrated variable and the notion of cointegration and common trends.

This lecture was presented at the University of Copenhagen for the occasion of Copenhagen as the Cultural City of Europe, September 1996, and at EUI for the workshop on Inflation and Unemployment in Economies in Transition, October 1996.

1 Introduction

The concept of cointegration was defined by Granger (1981) and after the paper by Engle and Granger (1987) it has become one of the cornerstones in modern time series econometrics, although it was implicitly applied by Sargan (1964) and Davidson, Hendry, Srba and Yeo (1978). It is the purpose of this paper to give some simple examples of cointegrated time series and discuss in relation to a very simple economic problem, how cointegration can be useful. Examples of mathematical statistical models which allow such phenomenon are given.

Section 2 gives a brief description of an economic problem and the data and in section 3 we discuss the notion of a mathematical model in particular a mathematical statistical model of economic data. Section 4 gives some examples of equations that can generate stationary and non-stationary univariate series and in section 5 the same is done for systems of variables that interact.

The notion of equilibrium (or error) correction model is introduced and the basic definition of integration, cointegration and common trends are given.

Throughout there are a few references to the existing literature.

2 A simple economic story and a data set

We give in this section an example of an economic problem and illustrate by a set of data which consists of quarterly measurements of prices and exchange rates from 1973 first quarter to 1987 third quarter. The data is shown in Figure 1. We shall use the data to illustrate the following simple economic problem.

The law of one price states that the price of a commodity in Denmark (P^{dk}) and the price of the same commodity (P^{ge}) in Germany are equal if expressed in the same currency. If ($E^{dk/ge}$) is the exchange rate between German Mark and Danish Krone then the law of one price is expressed as

$$P^{dk} = P^{ge}E^{dk/ge}. \tag{1}$$

If the prices measured are not prices of simple commodities but rather price indices computed in a roughly similar manner in different countries, one would expect that this law does not hold as an identity, but rather that if it does not hold over a longer period, then some adjustment must take place either in the exchange rate or in the prices, so that a disequilibrium between price levels will be diminished. If (1) holds we say that Purchasing Power Parity holds.

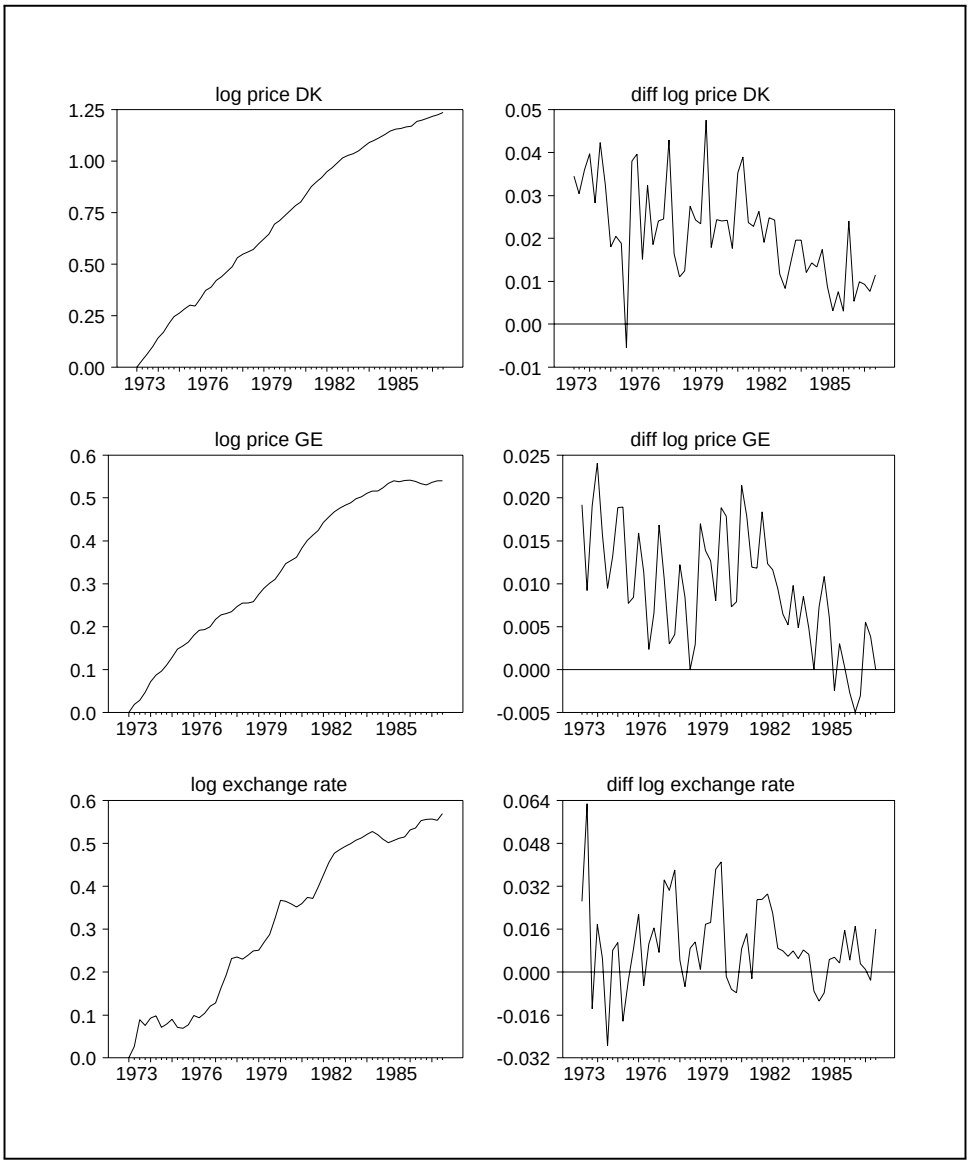

Figure 1: The data shown in the left hand panel are log price level in Denmark and Germany as well as the exchange rate. In the right hand panel we have shown the data in differences.

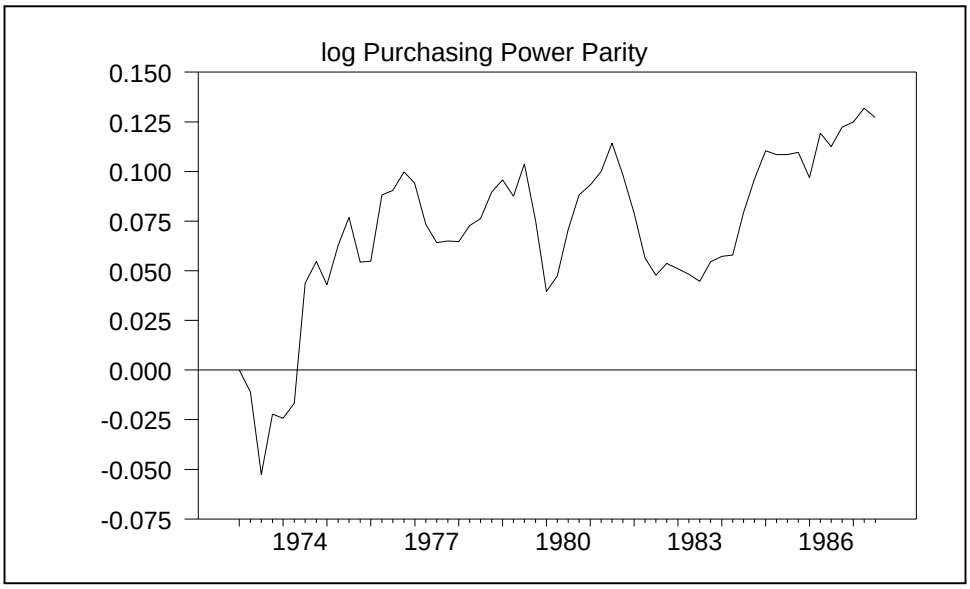

Figure 2: The series shown is $ppp_t = p_t^{dk} - p_t^{ge} - e_t$.

The data in Figure 1 show the consumer price series in Denmark and Germany and the exchange rate. As is usually done in this type of work we have expressed the variables in logarithms. The data is also presented in differences, such that the change of a price series in logs corresponds to a relative change in the price or in other words the inflation rate. Expressed in logarithms we get for $p^{dk} = \log(P^{dk})$, $p^{ge} = \log\left(P^{ge}\right)$, and $e = \log(E^{ge/dk})$ that the law one price states that

$$ppp_t = p_t^{dk} - p_t^{ge} - e_t = 0.$$

By plotting the variable ppp_t in Figure 2 it is seen that the law of one price does not hold with such a simple formulation. We note that the log prices increase roughly linearly corresponding to a constant inflation rate. The erratic movements around such a straight line is what we are concerned with in modelling by a statistical model. The exchange rate moves around in a seemingly random fashion and the differences, which form the quarterly increases, look a lot more stable, in the sense that the trend has been taken out of all three series.

The present paper deals with the problem of modelling time series as seen in Figures 1 and 2. It discusses some statistical concepts developed for the analysis of such models.

3 Modelling Economic Time series

The purpose of modelling economic variables is to describe the phenomenon we observe, but we hope by this description to get behind the immediate picture and express our understanding of the economic mechanism that governs the development.

The purpose of this can be prediction or simply understanding. If you understand the mechanism that produced the phenomenon, you can advise on policy implications and design policies for specific purposes.

It is a common feature of such models that they describe in some detail that part of reality which we observe, but not in every detail, since then it would be much too complicated. Hence a model constitutes a cleverly chosen simplification of reality.

Of special concern here are phenomenon that involve randomness, and which require statistical models.

It is an important aspect of a model that it can be incorrect, and it is in the confrontation between a theoretical model and observations of reality that statistics has a role to play.

When modelling the statistical fluctuations of economic variables, we can start with the idea of prediction. If we want to predict the exchange rate tomorrow on the basis of observations of past exchange rates, we define the prediction error as the difference between the observed exchange rate and our prediction of it.

The prediction error measures how much we missed the actual outcome and measures the surprise in the data.

Let e_t be the exchange rate measured in period t and let the past values observed be $e_1, \ldots, e_{t-1}$. These can be daily measurement, weekly measurements or, as on Figure 1, quarterly measurements. A simple, and probably too simple, prediction strategy is expressed in

$$\textit{prediction of exchange rate tomorrow} = \textit{exchange rate today.}$$

In this case the prediction error or surprise is just

$$e_t - e_{t-1}.$$

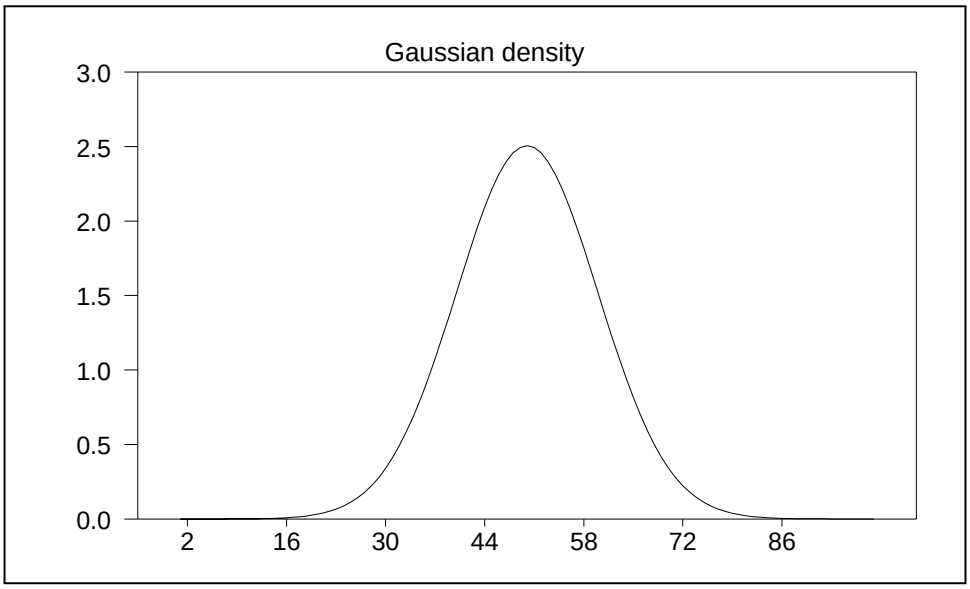

Figure 3: The Gaussian density with mean 50 and variance 100.

It has been the tradition since the work of Haavelmo (1944) to model surprises as the outcome of random experiments, or as random variables, and this is where statistical modelling comes in.

The simplest of such models is to define

$$\epsilon_t = e_t - e_{t-1}, \tag{2}$$

and then assume that the ϵ are independent random variables which follow the Gaussian distribution.

In Figure 4 the density of such a random variable is given. This density can be considered a continuous analogue of the binomial distribution in Figure 3. The interpretation of this curve is that the probability of finding the random variable between two numbers a and b is the area under the curve from a to b.

Such random numbers are easily generated on a computer and in Figure 5 we have shown a sequence of random Gaussian numbers.

The observations are connected by straight lines to emphasize the variation of the points around the line which gives the central value 0. Model (2) is clearly a very simple description of reality, but is sometimes surprisingly accurate. So much in fact that when econometricians want to model exchange rates they often compare their method of prediction, or their model, with the above simple benchmark.

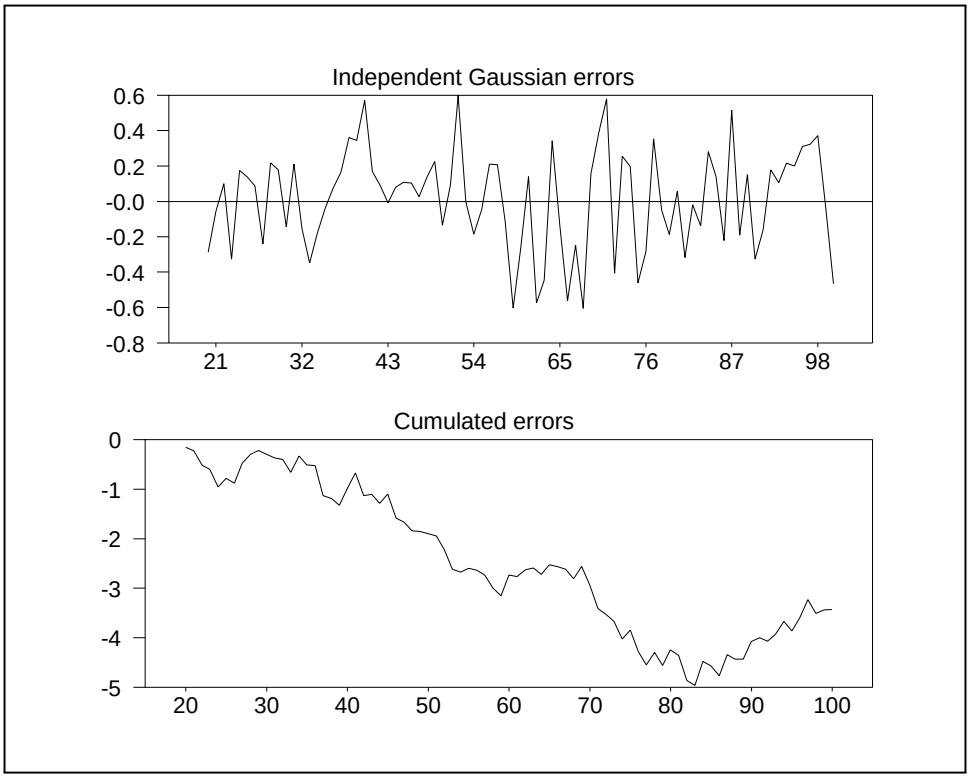

Figure 4: The left panel shows a sequence of independent Gaussian errors, and the in right hand panel they are cumulated to show a random walk.

The problem facing the modeler is how to extract information from past values concerning the predictions. We often model the prediction (or the deterministic part of the process) as a linear function of past values and let ϵ_t denote the un-modelled part of the process or the surprise.

3.1 Random shocks and intervention shocks

The above description of prediction errors as random is of course too simple. The economy often gives the econometrician surprises. Some of these can be quite easily modelled as random shocks, but others are clearly of more fundamental nature as the transition from a macro economic policy influenced by Keynesian economics to monetarist economics, or the shock that the economy suffered when the capital markets were liberalized around 1983, when capital was allowed to move almost freely across international boundaries. It would not be reasonable to try to neglect such serious disturbances in the economy, but rather analyse their influence on the variables in question. Thus shocks can be grouped into random shocks and intervention shocks. It is an important part of econometric model building to sort out which shocks are in which category.

4 Univariate time series

In order to illustrate the concepts used in modelling economic variables we first consider a single process e_t. We want to demonstrate how we can use equations with random disturbances to generate processes with desirable properties.

4.1 Non-stationary processes and the random walk

If we add equation (2) for $i = 1, \ldots, t$ we get the equation

$$e_t = e_0 + \epsilon_1 + \cdots + \epsilon_t. \tag{3}$$

Figure 5 shows such a process. We have here defined an example of a so-called stochastic process or time series. This particular one is called a random walk, a terminology which alludes to a person who staggers along a street taking steps (ϵ_t) of random size and direction, whereby e_t will be his position after t steps when he starts at e_0. The equations (2), when solved, gives a process which behaves very much like the exchange rate.

It should be emphasized, that by modelling the exchange rate by a random walk, we do not try to produce sample paths, or pictures, that look like the observed one in every detail, since we have decided that these details are unexplainable small shocks or surprises, which we want to model as random. But we want to use the model to create sample paths with the same qualitative behaviour, which loosely speaking can be described as a process that "floats" in the sense that once it has reached a level it stays there until it reaches a new level. Such a process is called non-stationary as opposed to a stationary process which we shall now discuss.

4.2 Autoregressive and stationary processes

Consider the disequilibrium error in the law of one price

$$ppp_t = p_{1t} - p_{2t} - e_{12t}.$$

Figure 2 shows that we have a process that is somewhat more stable than the prices and exchange rates, and which does not "float" as much as for instance the exchange rate. We want a model that mimics the adjusting behaviour of the prices and exchange rates, that is, a model with the property that if ppp_t gets too far away from zero, then it will be forced back towards zero either by changing prices or exchange rates.

This "reversal towards the mean" can be modelled by a so-called stationary process and a simple example of an equation that can generate such an adjusting process, which we shall call y_t, is

$$y_t - y_{t-1} = \pi(y_{t-1} - \mu) + \epsilon_t. \tag{4}$$

If for instance $\pi = -\frac{1}{2}$, and we have observed $y_{t-1} > \mu$, then the change in the process is minus one half times the overshoot $(y_{t-1} - \mu)$ plus a random noise, which means that if the process get too far out then a correction will take place and bring the process back towards the value μ. Figure 6 shows the simulations of a stationary process generated by (4) for two different values of the parameter π and $\mu = 0$. The processes generated by (4) are called autoregressive processes.

The name is chosen because in statistical jargon the model suggests a regression analysis of the value y_t on the past y_{t-1}.

Note that the value of π measures the "glue" in the process. It π is around -1, neighbouring values are almost unrelated or independent,a dn the process

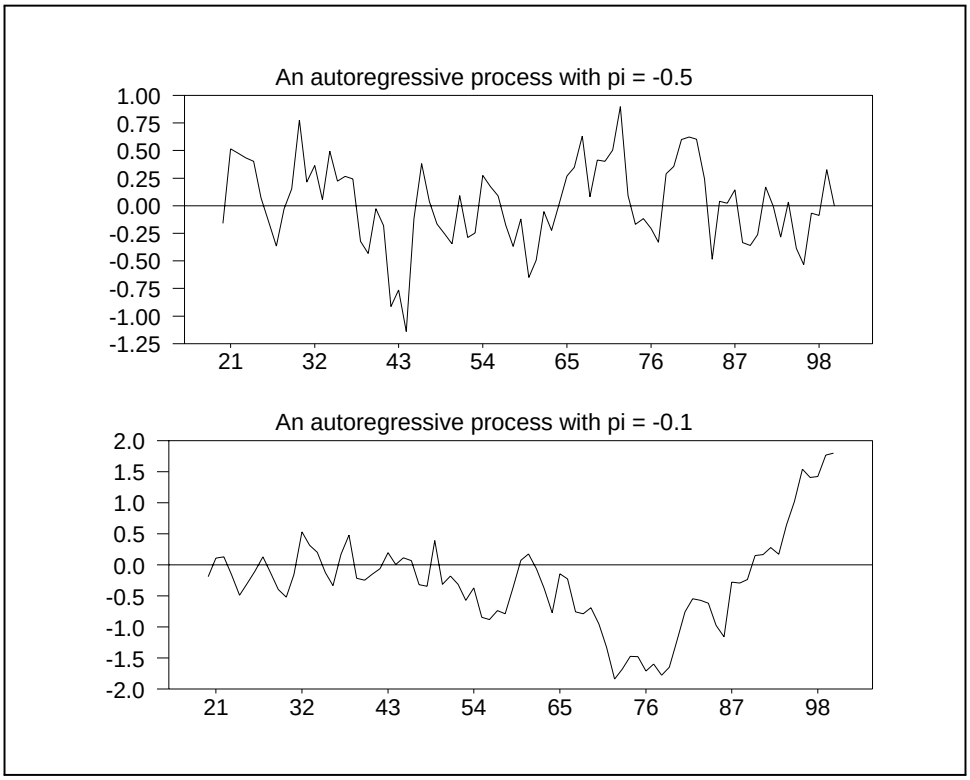

Figure 5: Autoregressive processes with parameters $\pi = -0.5$ and $\pi = -0.1$.

moves up and down completely irregularly. If π is close to 0 then neighbouring values are more often close together, and we get a wave-like behaviour.

Model (4) is an analogue of a differential equation, which in physics is used to model continuous dynamic phenomena. The simplest differential equation is

$$\frac{dy(t)}{dt} = \pi(y(t) - \mu), \tag{5}$$

which has solution

$$y(t) = (y(0) - \mu)e^{\pi t} + \mu.$$

Thus for $\pi < 0$ we get an exponential decay towards μ.

Model (4) is a simplification of (5) in the sense that we only have observations at distinct time points, but more complicated than (5) since we also have random disturbances.

In summary we can say that model (4) generates a process y_t which, if $-2 < \pi < 0$, exhibits mean reversion or stationarity, and if $\pi = 0$ we get a random walk, which is non-stationary.

Note how the deterministic part of the change of the process is modelled by $\pi y_{t-1} + \mu$, and the random part by the Gaussian random number ϵ_t.

5 Multivariate systems

The above discussion of stationarity and non-stationarity focuses on a single variable observed over time. In practice we often have many variables observed together and we here discuss such systems, and use that as a starting point for defining the important notions of equilibrium correction models, integration and cointegration.

Consider therefore the system consisting of the three variables from section 2. We want to model these by a multivariate version of the autoregressive process considered in (4). We suggest the model

$$\begin{aligned} p_t^{dk} - p_{t-1}^{dk} &= \pi_{11}p_{t-1}^{dk} + \pi_{12}p_{t-1}^{ge} + \pi_{13}e_{t-1} + \mu_1 + \epsilon_{1t}, \\ p_t^{ge} - p_{t-1}^{ge} &= \pi_{21}p_{t-1}^{dk} + \pi_{22}p_{t-1}^{ge} + \pi_{23}e_{t-1} + \mu_2 + \epsilon_{2t}, \\ e_t - e_{t-1} &= \pi_{31}p_{t-1}^{dk} + \pi_{32}p_{t-1}^{ge} + \pi_{33}e_{t-1} + \mu_3 + \epsilon_{3t}. \end{aligned} \tag{6}$$

Expresses in words the changes of each of the variables is predicted by a linear combination of past values of all variables. One can make this a lot more

complicated by adding more past values or even other variables which can help explain the variability in the prices and exchange rate. For now (6) is enough to illustrate the concepts needed in the empirical analysis of economic data.

The coefficients π_{ij} measure the effect that each of the past values have on the changes, and μ_i is related to the growth rates.

5.1 The equilibrium or error correction model

The first question that one can ask one self is: What are the properties of the processes p_t^{dk}, p_t^{ge} and e_t generated by such equations. We want to be able to generate processes that have qualitatively the same properties as the observed series.

The answer to this problem is of course that it depends on which parameter values we consider.

It turns out that for the equations (6) to generate non-stationary variables we need a condition that mimics the condition $\pi = 0$ in the univariate situation. The algebra behind this is difficult, but suffice it here to say that if the coefficients π_{ij} have the special structure given by the equations

$$\begin{aligned} p_t^{dk} - p_{t-1}^{dk} &= \alpha_1(p_{t-1}^{dk} - p_{t-1}^{ge} - e_{t-1}) + \epsilon_{1t}, \\ p_t^{ge} - p_{t-1}^{ge} &= \alpha_2(p_{t-1}^{dk} - p_{t-1}^{ge} - e_{t-1}) + \epsilon_{2t}, \\ e_t - e_{t-1} &= \alpha_3(p_{t-1}^{dk} - p_{t-1}^{ge} - e_{t-1}) + \epsilon_{3t}, \end{aligned} \tag{7}$$

then the processes have a number of desirable properties.

This is a special case of the general equation (6), with $\pi_{11} = \alpha_1, \pi_{12} = -\alpha_1$,$\pi_{13} = -\alpha_1$, etc. Note that the only linear combination of the past values that enters is the ppp_{t-1}. Thus the past speaks to us through the value of the equilibrium error.

The interpretation of these equations is that the change in the prices, and exchange rates react through the coefficients α_i to a deviation between the prices and exchange rates as measured by the ppp_{t-1} in the past. Thus if the Danish prices are too large for some period, relative to the German prices as measured by the ppp_{t-1} then the Danish prices will react to this disequilibrium and decrease. In this case it is not so obvious that the German prices will react, because their economy is so much larger, so we allow the value 0 for the coefficient α_2. Finally the coefficient α_3 measures the reaction of the exchange rates to the high prices in Denmark.

This simple story shows how we can capture a backward looking adjustment behaviour in the framework of a simple tri-variate stochastic difference equation. It should be obvious that we are assuming linearity, whereas a non-linear reaction function could be more realistic. It is also quite obvious that the formation of exchange rates is related to the financial markets, hence the interest rates would be relevant variables to enter into the model, and the data. Thus there is ample room for making the modelling more realistic. This, however, is not the intention of the present presentation.

5.2 Integration, cointegration and common trends

We discuss in this section three important concepts that have made the modelling of random walk type phenomenon possible.

First of all one can show that the processes generated by (6) are non-stationary, that is, they do not have any tendency to revert to a mean but "float", like random walks, see Figure 3. We say that such processes are *integrated*, a terminology which comes from the usual notation of an integral through the following associations: An integral $\int$ is the limit of a sum, but a random walk is exactly a sum of the small shocks. Hence we cumulate or sum or integrate the shocks to get a random walk which is an integrated process.

Moreover, we can show that the linear combination

$$ppp_t = p_t^{dk} - p_t^{ge} - e_t,$$

is mean reverting or stationary. We say that the processes p_t^{dk}, p_t^{ge} and e_t are *cointegrated.* We have simulated a bivariate system of cointegrated variables in Figure 5.

Note how the individual processes float, whereas the difference between the processes as measured by ppp_t is stationary, such that the processes move together. It is this co-movement phenomenon that we want to use in modelling economic variables in the following way.

Often economic theory makes *static* statements about relations between economic variables, which in reality move in a *dynamic* fashion. How can one reconcile a static statement as the law of one price

$$p_{1t} - p_{2t} - e_{12t} = 0,$$

with the dynamically evolving data as shown in Figure 2 ?

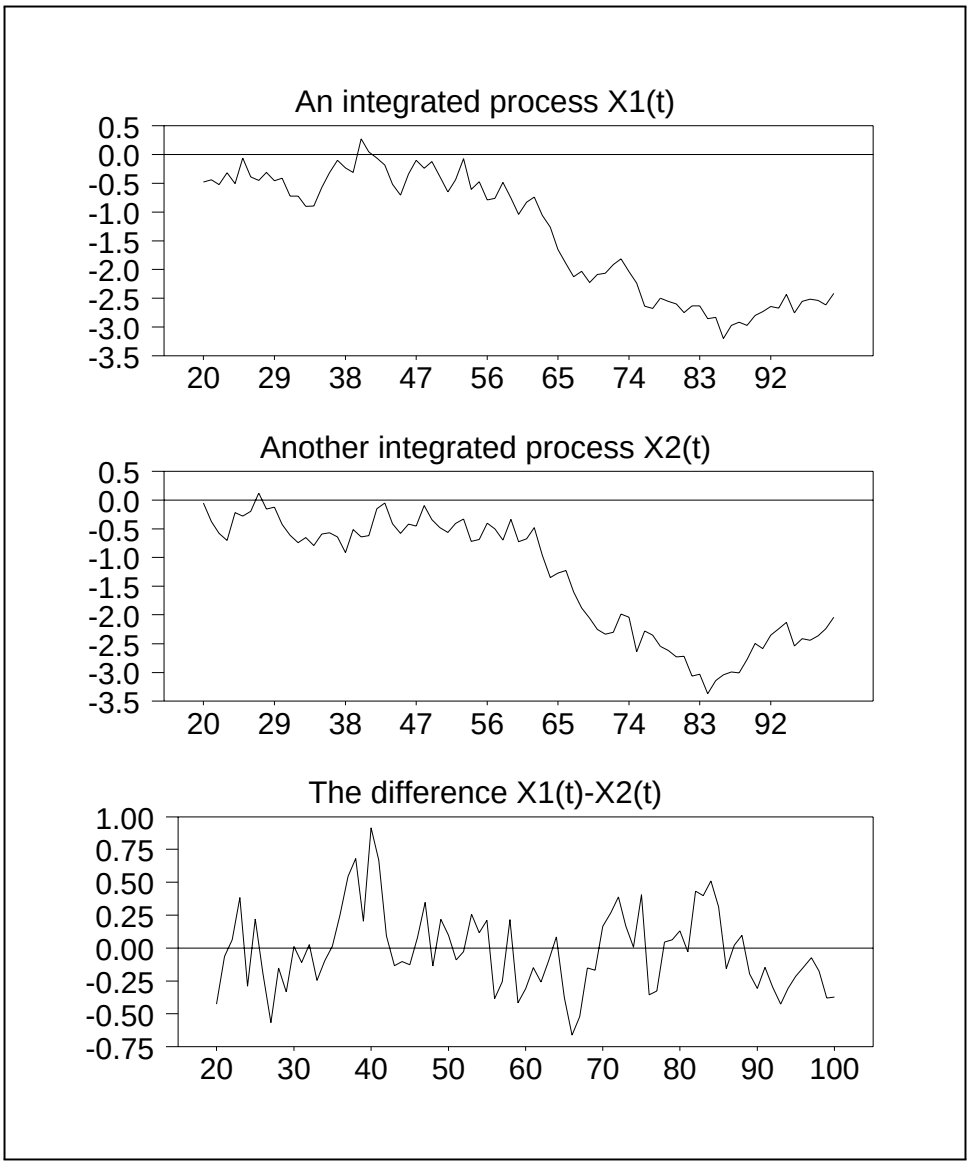

Figure 6: The figure shows two processes generated from the equations $X_{1t} - X_{1t-1} = -\frac{1}{4}(X_{1t-1} - X_{2t-1}) + \epsilon_{1t}$, and $X_{2t} - X_{2t-1} = \frac{1}{4}(X_{1t-1} - X_{2t-1}) + \epsilon_{2t}$. Note that both processes are integrated but their difference is stationary, hence X_{1t} and X_{2t} are cointegrated.

Clearly "we need a stochastical formulation to make simplified relations elastic enough for applications", Haavelmo (1943), and the interpretation of the long-run relation expressed by PPP is in the present framework that ppp_t is stationary.

Thus, if an error correction model of the type (6) can describe the variation of the data, then we have achieved a description that allows individual variables to be non-stationary, as they appear to be in reality, and still have the property that a linear relation between the variables is stationary. Thus a cointegrating relation can be considered the statistical formulation of a long-run relation in the economy. Note that a long-run relation is something that the processes would satisfy if all errors were switched off and the equations would then bring back the processes to a set of values where the relation is satisfied, a steady state.

Another interpretation is that a long-run relation is like the equilibrium position of a pendulum. It is usually never seen in practice for a system that works, but can be considered an equilibrium position that is always present and towards which the pendulum is pulled.

5.3 The general ECM

We shall meet many examples of equilibrium correction models in the lectures by Juselius and Hendry, so let me just give the full formulation of the general error correction model expressed in mathematical form

$$\Delta X_t = \alpha\beta' X_{t-1} + \Gamma\Delta X_{t-1} + \mu + \Phi D_t + \epsilon_t. \tag{8}$$

The process $X_t = (X_{1t}, \ldots, X_{pt})'$ contains the relevant set of variables, often called the information set. The coefficients in the matrix α are the adjustment coefficients and the columns of β are the cointegrating vectors and $\beta' X_{t-1} = c$, the long-run relations. We call Γ the short-run coefficients, since they express how the changes react to past changes. Finally we allow for the processes to have a growth rate μ and include the term D_t to allow for extra variables like seasonal dummies or intervention dummies, that need to be taken into account when modelling the data.

Thus the general ECM allows us to model processes that are non-stationary but have stationary linear relations. The processes are composed of deterministic trends, of common stochastic trends or random walks and are called integrated. Cointegration appears when there are fewer common stochastic trends

than variables, such that these are driven by the same trends. This implies that they can be eliminated by suitable linear combinations, and that is how the phenomenon of cointegration appears as a useful way of modeling the long-run economic relations, that exhibit stability in a system of variables that all evolve dynamically.

Equilibrium correction models have been applied to describe economic data series since the work of Sargan (1964) and gained popularity with the paper by Davidson, Hendry, Srba and Yeo (1978), but it was the papers by Granger (1981) and in particular Engle and Granger (1987) that clarified the relation between ECM and cointegration, in the sense that any ECM will generate cointegrated variables and cointegrated variables can be expresses as solutions of equilibrium correction models.

5.4 The purpose of econometric modelling

The purpose of this type of econometric analysis can now briefly be formulated as follows. We have a set of economic variables which have been chosen because they represent measurements of interest to the economist. We have available theory of how these variables interact and are related, often expressed in terms of linear relations. We then analyse the data to find a reasonable statistical description, which in this context is an error correction model, in which we can formulate the economic theory as restrictions on the parameters, in particular we can often formulate the economic relations, that are derived for en economy in equilibrium or steady state, as the long-run or cointegrating relations.

The mathematical statistical analysis of such models amounts to checking whether error correction models provide an adequate description of the observed data. In the process of doing this, one will have to find reasonable estimators of all the unknown parameters and give guidelines for testing hypotheses on the parameters. Thus establishing the tools for checking the economic theory against reality. The book by Hendry (1995) is a general overview of econometric modelling of dynamic phenomena, and the paper by Johansen and Juselius (1990) gives the mathematical statistical analysis of the ECM with respect to inference on cointegration. A systematic theory of cointegration in the ECM is given in Johansen (1995).

6 References

Davidson, J. E. H., Hendry, D. F., Srba F. and Yeo S. (1978), Economic Modelling of the Aggregate Time Series Relationships between Consumers' Expenditure and Income in the United Kingdom, *Economic Journal,* 88, 661-692.

Engle, R. F. and Granger C. W. J. (1987), Cointegration and Error Correction: Representation, Estimation and Testing, *Econometrica*, 55, 251-276.

Granger C. W. J. (1981), Some Properties of Time Series and their use in Econometric Model Specification, *Journal of Econometrics*, 16, 121-130.

Haavelmo, T. (1943), The Statistical Implications of a System of Simultaneous equations, *Econometrica*, 11, 1-12.

Haavelmo, T. (1944), The Probability Approach in Econometrics, *Econometrica*, 12, 1-118. Supplement.

Hendry, D. F. (1985), *Dynamic Econometrics,* Oxford University Press, Oxford.

Johansen, S. (1995), *Likelihood-based inference in cointegrated vector autoregressive models,* Oxford University Press, Oxford.

Johansen, S. and Juselius, K. (1990), Maximum Likelihood Estimation and Inference on Cointegration with Applications to the Demand for Money, *Oxford Bulletin of Economics and Statistics,* 52, 169-210.

Sargan. J. D. (1964), Wages and Prices in the United Kingdom: A Study in Econometric Methodology (with discussion). In Hart, P. E. and Mills, G. and Whitaker, J. K. (eds.), *Econometric analysis for National Economic Planning*, 16, 25-63.